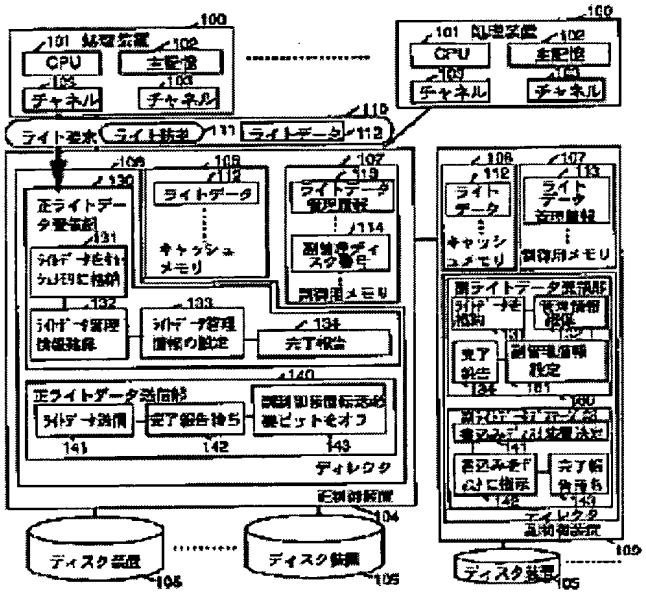
STORAGE CONTROL DEVICE

Patent number: JP11085408
Publication date: 1999-03-30
Inventor: YAMAMOTO AKIRA; NAKAMURA KATSUNORI; KIJIRO SHIGERU
Applicant: HITACHI LTD
Classification:
- international: G06F3/06; G06F3/06
- european:
Application number: JP19970248177 19970912
Priority number(s): JP19970248177 19970912

Also published as:
E P0902370 (A2)
US 6408370 (B2)
US 2001029570 (A1)

Abstract of **JP11085408**

PROBLEM TO BE SOLVED: To provide a function capable of minimizing the number of data transfers at the time of executing double writing between remote control devices, suppressing the deterioration of performance to the minimum even when the distance between the control devices is enlarged and preventing the halfway result of a transaction from being left, also to make it unnecessary to execute disk I/O processing for control information, and to attain high performance. SOLUTION: After returning write data completion report, a master control device 104 sends the reports directly to a sub-control device 109. The device 109 stores the received write data in a non-volatile memory to guarantee the data. Then, a time to be a certain reference is set up, all write data before the time are guaranteed and all write data after the time are discarded.



(19)日本国特許庁（ＪＰ）　(12)**公開特許公報**（Ａ）　(11)特許出願公開番号

特開平11－85408

(43)公開日　平成11年(1999)３月30日

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｇ０６Ｆ　3/06	３０１	Ｃ０６Ｆ　3/06	３０１Ｘ
	３０４		３０４Ｅ

審査請求　未請求　請求項の数８　ＯＬ　（全 14 頁）

(21)出願番号　特願平9－248177

(22)出願日　平成９年(1997)９月12日

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地

(72)発明者　山本　彰
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内

(72)発明者　中村　勝彦
神奈川県小田原市国府津2880番地　株式会社日立製作所ストレージシステム事業部内

(72)発明者　木城　茂
神奈川県小田原市国府津2880番地　株式会社日立製作所ストレージシステム事業部内

(74)代理人　弁理士　小川　勝男

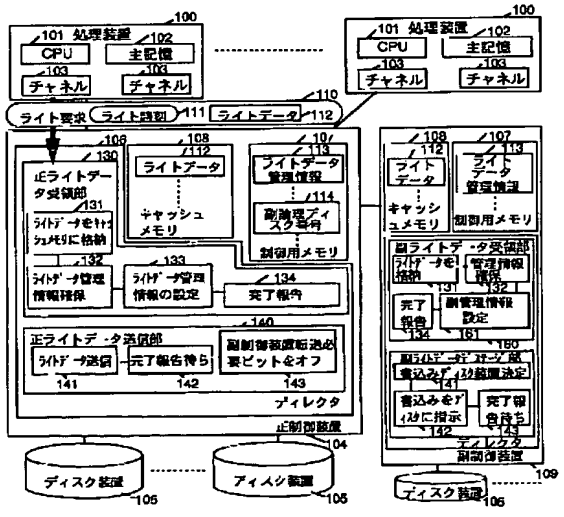
(54)【発明の名称】　記憶制御装置

(57)【要約】

【課題】本発明の課題は、遠隔地の制御装置間で２重書きを行う際、データ転送数を必要最小限に留め、制御御装置間の距離が拡大しても、性能の劣化を微小に押さえ、さらに、トランザクションの途中結果を残さないような機能を提供することである。加えて、制御情報のディスク入出力処理の実行も不必要とし、高性能化を図る点にある。

【解決手段】本発明では、正制御装置は、ライトデータを完了報告を返した後、直接副制御装置に送る。さらに、副制御装置は、受け取ったライトデータを不揮発メモリに格納することで、データ保証を行う。さらに、ある基準となる時刻を設け、この時刻以前のすべてのライトデータを保証し、この時刻より後のライトデータはすべて破棄できるようにする。

図１



【特許請求の範囲】
【請求項１】それぞれの制御装置が１台以上の記憶装置を接続し、不揮発メモリを有する複数の制御装置から構成される複合記憶装置システムであって、
前記複数の制御装置の内の１台以上の制御装置が、
処理装置から、ライト要求を受付、前記ライト要求は前記ライト要求が発行されたライト時刻を含み、前記ライト要求において転送されたライトデータを前記不揮発メモリに格納する手段と、
前記処理装置に、前記ライト要求の完了報告を行う手段と、
前記複合記憶システムを構成する他の制御装置に、前記ライトデータと前記ライト時刻を送る手段とを有し、
前記複数の制御装置の内の１台以上の制御装置が、
前記複合記憶システムを構成する他の制御装置から、前記ライトデータと前記ライト時刻を受け取り、前記不揮発メモリに格納する手段とを有し、
前記ライト時刻を参照して、前記ライトデータを前記記憶装置に書き込む手段を有することを特徴とする複合記憶装置システム。
【請求項２】それぞれの制御装置が１台以上の記憶装置を接続し、不揮発メモリを有する複数の制御装置から構成される複合記憶装置システムであって、
前記複数の制御装置の内の１台以上の制御装置が、
処理装置から、ライト要求を受付、前記ライト要求は前記ライト要求が発行されたライト時刻を含み、前記ライト要求において転送されたライトデータを前記不揮発メモリに格納する手段と、
前記処理装置に、前記ライト要求の完了報告を行う手段と、
前記複合記憶システムを構成する他の制御装置に、前記ライトデータと前記ライト時刻を送る手段とを有し、
前記複数の制御装置の内の１台以上の制御装置が、
前記複合記憶システムを構成する他の制御装置から、前記ライトデータと前記ライト時刻を受け取り、前記不揮発メモリに格納する手段とを有し、
前記ライト時刻を参照して、前記ライトデータを前記記憶装置に書き込む手段を有し、
前記ライト時刻を参照して、前記ライトデータを前記不揮発メモリから破棄する手段を有することを特徴とする複合記憶装置システム。
【請求項３】それぞれの制御装置が１台以上の記憶装置を接続し、不揮発メモリを有する２台の制御装置から構成される複合記憶装置システムであって、
前記２台の制御装置の内の１台の制御装置が、
処理装置から、ライト要求を受付、前記ライト要求は前記ライト要求が発行されたライト時刻を含み、前記ライト要求において転送されたライトデータを前記不揮発メモリに格納する手段と、
前記処理装置に、前記ライト要求の完了報告を行う手段と、
前記複合記憶システムを構成する他の制御装置に、前記ライト時刻順に前記ライトデータを送る手段とを有し、
前記２台の制御装置の内の１台の制御装置が、
前記複合記憶システムを構成する他の制御装置から、前記ライトデータを受け取り、前記不揮発メモリに格納する手段とを有し、
前記ライトデータを前記記憶装置に書き込む手段を有することを特徴とする複合記憶装置システム。
【請求項４】それぞれの制御装置が１台以上の記憶装置を接続し、不揮発メモリを有する複数の制御装置から構成される複合記憶装置システムであって、
前記複数の制御装置の内の１台以上の制御装置が、
処理装置から、ライト要求を受付、前記ライト要求は前記ライト要求が発行されたライト時刻を含み、前記ライト要求において転送されたライトデータを前記不揮発メモリに格納する手段と、
前記処理装置に、前記ライト要求の完了報告を行う手段と、
前記複合記憶システムを構成する他の制御装置に、前記ライトデータと前記ライト時刻を送り、前記他の制御装置から前記ライトデータと前記ライト時刻の受領報告を受け取る手段と前記複合記憶システムを構成する前記他の制御装置に、前記他の制御装置から、受領報告を受け取った前記ライトデータに関するライト時刻である受領報告ライト時刻を送る手段を有し、
前記複数の制御装置の内の１台以上の制御装置が、
前記複合記憶システムを構成する他の制御装置から、前記ライトデータと前記ライト時刻を受け取り、前記不揮発メモリに格納し、前記前記ライトデータと前記ライト時刻の受領報告を、前記他の制御装置に送る手段とを有し、
前記複合記憶システムを構成する前記他の制御装置から、前記他の制御装置から、前記受領報告ライト時刻を受け取る手段を有し、
前記ライト時刻と前記受領報告ライト時刻を参照して、前記ライトデータを前記記憶装置に書き込む手段を有することを特徴とする複合記憶装置システム。
【請求項５】それぞれの制御装置が１台以上の記憶装置を接続し、不揮発メモリを有する複数の制御装置から構成される複合記憶装置システムであって、
前記複数の制御装置の内の１台以上の制御装置が、
処理装置から、ライト要求を受付、前記ライト要求は前記ライト要求が発行されたライト時刻を含み、前記ライト要求において転送されたライトデータを前記不揮発メモリに格納する手段と、
前記処理装置に、前記ライト要求の完了報告を行う手段と、
前記複合記憶システムを構成する他の制御装置に、前記ライトデータと前記ライト時刻を送り、前記他の制御装

置から前記ライトデータと前記ライト時刻の受領報告を受け取る手段と前記複合記憶システムを構成する前記他の制御装置に、前記他の制御装置から、受領報告を受け取った前記ライトデータに関するライト時刻である受領報告ライト時刻を送る手段を有し、
前記複数の制御装置の内の１台以上の制御装置が、
前記複合記憶システムを構成する他の制御装置から、前記ライトデータと前記ライト時刻を受け取り、前記不揮発メモリに格納し、前記前記ライトデータと前記ライト時刻の受領報告を、前記他の制御装置に送る手段とを有し、
前記複合記憶システムを構成する前記他の制御装置から、前記他の制御装置から、受領報告ライト時刻を受け取る手段を有し、
前記ライト時刻をと前記受領報告ライト時刻を参照して、前記ライトデータを前記記憶装置に書き込む手段と、
前記ライト時刻と前記受領報告ライト時刻を参照して、前記不揮発性メモリから、前記ライトデータを消去する手段を有することを特徴とする複合記憶装置システム。
【請求項６】それぞれの制御装置が１台以上の記憶装置を接続し、不揮発メモリを有する複数の制御装置から構成される複合記憶装置システムであって、
前記複数の制御装置の内の１台以上の制御装置が、
処理装置から、ライト要求を受付、前記ライト要求は前記ライト要求が発行されたライト時刻を含み、前記ライト要求において転送されたライトデータを前記不揮発メモリに格納する手段と、
前記処理装置に、前記ライト要求の完了報告を行う手段と、
前記複合記憶システムを構成する他の制御装置に、前記ライトデータと前記ライト時刻を送る手段とを有し、
前記複数の制御装置の内の１台以上の制御装置が、
前記複合記憶システムを構成する他の制御装置から、前記ライトデータと前記ライト時刻を受け取り、前記不揮発メモリに格納する手段とを有し、
受け取った前記ライト時刻に関する情報を他の制御装置に送る手段と、
前記ライト時刻を参照して、前記ライトデータを前記記憶装置に書き込む手段を有することを特徴とする複合記憶装置システム。
【請求項７】それぞれの制御装置が１台以上の記憶装置を接続し、不揮発メモリを有する複数の制御装置から構成される複合記憶装置システムであって、
前記複数の制御装置の内の１台以上の制御装置が、
処理装置から、ライト要求を受付、前記ライト要求は前記ライト要求が発行されたライト時刻を含み、前記ライト要求において転送されたライトデータを前記不揮発メモリに格納する手段と、
前記処理装置に、前記ライト要求の完了報告を行う手段と、
前記複合記憶システムを構成する他の制御装置に、前記ライトデータと前記ライト時刻を送り、前記他の制御装置から前記ライトデータと前記ライト時刻の受領報告を受け取る手段と前記複合記憶システムを構成する前記他の制御装置に、前記他の制御装置から、受領報告を受け取った前記ライトデータに関するライト時刻である受領報告ライト時刻を送る手段を有し、
前記複数の制御装置の内の１台以上の制御装置が、
前記複合記憶システムを構成する他の制御装置から、前記ライトデータと前記ライト時刻を受け取り、前記不揮発メモリに格納し、前記前記ライトデータと前記ライト時刻の受領報告を、前記他の制御装置に送る手段とを有し、
前記複合記憶システムを構成する前記他の制御装置から、前記他の制御装置から、前記受領報告ライト時刻を受け取る手段を有し、
受け取った前記受領報告ライト時刻に関する情報を、他の制御装置に送る手段と、
前記ライト時刻を参照して、前記ライトデータを前記記憶装置に書き込む手段を有することを特徴とする複合記憶装置システム。
【請求項８】それぞれの制御装置が１台以上の記憶装置を接続し、不揮発メモリを有する複数の制御装置から構成される複合記憶装置システムであって、
前記複数の制御装置の内の１台以上の制御装置が、
処理装置から、ライト要求を受付、前記ライト要求は前記ライト要求が発行されたライト時刻を含み、前記ライト要求において転送されたライトデータを前記不揮発メモリに格納する手段と、
前記処理装置に、前記ライト要求の完了報告を行う手段と、
前記複合記憶システムを構成する他の制御装置に、前記ライトデータと前記ライト時刻を送り、前記他の制御装置から前記ライトデータと前記ライト時刻の受領報告を受け取る手段と前記複合記憶システムを構成する前記他の制御装置に、前記他の制御装置から、受領報告を受け取った前記ライトデータに関するライト時刻である受領報告ライト時刻を送る手段を有し、
前記複数の制御装置の内の１台以上の制御装置が、
前記複合記憶システムを構成する他の制御装置から、前記ライトデータと前記ライト時刻を受け取り、前記不揮発メモリに格納し、前記前記ライトデータと前記ライト時刻の受領報告を、前記他の制御装置に送る手段とを有し、
前記複合記憶システムを構成する前記他の制御装置から、前記他の制御装置から、受領報告ライト時刻を受け取る手段を有し、
受け取った前記受領報告ライト時刻に関する情報を、他の制御装置に送る手段と、

前記ライト時刻を参照して、前記ライトデータを前記記憶装置に書き込む手段と、
前記ライト時刻を参照して、前記不揮発性メモリから、前記ライトデータを消去する手段を有することを特徴とする複合記憶装置システム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、異なった制御装置の間で、データを２重書きする機能に関する。特に、制御装置間の距離が長く、制御装置間のデータ転送に遅延が発生するような場合、本発明は、有効である。
【０００２】
【従来の技術】本発明に関する公知例として、以下の技術が開示されている。
【０００３】European Patent Application publication number ０６７１６８６Ａ１は、では、遠隔地にある制御装置間のディスクの２重書きを行う技術が、開示されている。本発明では、一方の制御装置が、地震などの天災等により破壊されても、もう一方の制御装置のディスクでデータ保証が可能というものである。European Patent Application publication number ０６７１６８６Ａ１では、ホスト計算機から、直接、ライトデータを受領する正側の制御装置は、遠隔地にある副側の制御装置へ、受領したライトデータを転送した後、ライトデータの受領完了を、ホスト計算機に報告する。この方法であると、正側と副側で完全にデータが一致するため、データ保証の点からは、非常に良い方法であった。しかし、制御装置間の距離拡大により、制御装置の間のデータ転送時間は、非常に大きくなるため、遠距離時に、性能上の課題があった。
【０００４】European Patent Application publication number ０６７２９８５Ａ１でも、遠隔地にある制御装置間のディスクの２重書きを行う技術が、開示されている。European Patent Application publication number ０６７２９８５Ａ１では、正側のホスト計算機から、直接、ライトデータを受領する正側の制御装置は、ライトデータ受領後直ちに、ライトデータの受領完了を、正側のホスト計算機に報告する。European Patent Application publication number ０６７２９８５Ａ１では、さらに、正側の制御装置が受領したライトデータのコピーが、一度正側のホスト計算機に読みだされる。本発明では、当初正側のホスト計算機から受領するライトデータには、時刻が付与されている。時刻は、このライトデータを書き込むライト要求が発行された時刻を意味する。ライトデータのコピーが、正側のホスト計算機に読みだされる時、ライト時刻も正側のホスト計算機に渡される。この後、正側のホスト計算機はライトデータのコピーとライト時刻を、副側のホスト計算機に送る。
【０００５】ライトデータとライト時刻を受け取ったを副側のホスト計算機は、ライト時刻等の情報を、制御用のディスクに書き込む。さらに、各ライトデータに付与された時刻を参照し、ライト時刻順に、ライトデータを副側のディスクに書き込みを行う。
【０００６】European Patent Application publication number ０６７２９８５Ａ１で副側のホスト計算機が、上記のような処理を行う目的は、オンラインシステム等で標準的に使用されるトランザクションの途中結果を残さないようにするためである。例えば、口座Ａから口座Ｂに預金を移すトランザクションを実行する場合、口座Ａから預金を引き落としたにもかかわらず、口座Ｂに預金を積み立てない状態を残さないようにすることが、トランザクションの途中結果を残さないということを意味する。通常、オンラインシステムでは、回復の単位は、トランザクションであるため、トランザクションの途中結果を残すことは、極めて重要な障害である。
【０００７】次に、上記のような処理を実行すると、トランザクションの途中結果を残さないようにすることができることを簡単に説明する。２重書きを行っているディスクの中には、口座情報等のデータベースを格納したディスクと、トランザクションの更新履歴を残したジャーナルを格納したディスクがある。ホスト計算機がダウンすると、回復処理プログラムにより、ジャーナルが解析され、終了していないトランザクションの更新結果は、実行開始前の状態に戻される等の処理が、実行され、トランザクションの途中結果を残さないようにすることができる。副側の制御装置のディスクに書き込んだライトデータが、有効となるのは、最新のライトデータを格納した正側の制御装置が破壊されてしまったような場合である。副側の制御装置には、最新のライトデータは格納されてはいないが、ある時刻までのライトデータは保証されていることになる。したがって、見かけ上、ホスト計算機が、ライトデータを保証している時刻に、ダウンしたのと等価な状態を作り出していることになる。したがって、副制御装置側のジャーナルを格納したディスクと、データベースを格納したディスクを用いて、ホスト計算機がダウンした時に実行される回復処理と同様の処理を実行することにより、トランザクションの途中結果を残さないようにすることができる。
【０００８】特開平４－２４５３４２は、ディスク制御装置が不揮発性のキャッシュメモリを持ち、ライトアフタを行う、すなわち、ホスト計算機から受領したライトデータを不揮発性のキャッシュメモリに書き込み、完了報告を行う技術が開示されている。不揮発性のキャッシュメモリは、信頼性が高いため、ここにライトデータを

格納すれば、十分データ保証が可能となると判断できるためである。

【０００９】

【発明が解決しようとする課題】Ｅｕｒｏｐｅａｎ Ｐａｔｅｎｔ Ａｐｐｌｉｃａｔｉｏｎ ｐｕｂｌｉｃａｔｉｏｎ ｎｕｍｂｅｒ ０６７２９８５Ａ１では、制御装置間の距離が拡大しても、若干のデータは、失われるものの、性能の劣化は、少ない。しかも、トランザクションの途中結果を残さない。

【００１０】しかし、正側のホスト計算機がデータを読みだし、副側のホスト計算機にデータを転送するため、Ｅｕｒｏｐｅａｎ Ｐａｔｅｎｔ Ａｐｐｌｉｃａｔｉｏｎｐｕｂｌｉｃａｔｉｏｎ ｎｕｍｂｅｒ ０６７１６８６Ａ１のように、直接制御装置間で、ライトデータを受け渡す場合に比較し、データ転送が一度余分に実行される。さらに、ＭＴなどの記憶媒体への入出力処理の実行も必要となる。

【００１１】本発明の目的は、Ｅｕｒｏｐｅａｎ Ｐａｔｅｎｔ Ａｐｐｌｉｃａｔｉｏｎｐｕｂｌｉｃａｔｉｏｎ ｎｕｍｂｅｒ ０６７２９８５Ａ１のように、直接制御装置間で、ライトデータを受け渡し、しかも、制御装置間の距離が拡大しても、性能の劣化を微小に押さえ、しかも、トランザクションの途中結果を残さないような機能を提供することである。さらに、制御情報などのディスクへの入出力処理の実行も不必要とし、高性能化を図る。

【００１２】

【課題を解決するための手段】以下、本発明が、以上述べてきた目的をいかに実現するかについて述べる。

【００１３】本発明では、ホスト計算機は、正制御装置に、ライト要求を発行する際に、ライトデータにライト時刻を付与する。正制御装置は、ライトデータをホスト計算機から受け取ると、完了を報告する。この後、正制御装置は、副制御装置に、ライトデータとライト時刻を送る。この時、正の制御装置は、ライト時刻順に、ライトデータを、副の制御装置に送る。以上により、制御装置間の距離が拡大しても、性能の劣化を微小に押さえることができる。

【００１４】副側の制御装置では、正側の制御装置から受け取ったライトデータとを、不揮発のキャッシュメモリに格納する。これにより、制御情報などのディスクへの入出力処理なしに、ライトデータのデータ保証が可能となる。

【００１５】副側の制御装置では、受け取ったライト時刻を参照して、ある時刻までのライトデータを保証するようにする。これにより、トランザクションの途中結果を残さないようにすることが可能である。

【００１６】

【発明の実施の形態】以下、本発明の実施例を説明する。まず、第１の実施例について説明する。

【００１７】図１は、第１の実施例の概要を表す。第１の実施例における構成は、１台以上の処理装置１００、１台の正制御装置１０４、正制御装置１０４に接続された１台以上のディスク装置１０５、１台の副制御装置１０９、副制御装置１０９に接続された１台以上のディスク装置１０５より構成する。処理装置１００は、ＣＰＵ１０１、主記憶１０２、チャネル１０３から構成される場合もある。正制御装置１０４は、制御用メモリ１０７、キャッシュメモリ１０８を含む。制御用メモリ１０７、キャッシュメモリ１０８は、不揮発化されているものとする。また、さらなる高信頼化のために、それぞれが２重化されていてもよい。キャッシュメモリ１０８、制御用メモリ１０７は、半導体メモリで構成されており、ディスク装置１０５に比べ、１桁から２桁高速なアクセスが可能である。正制御装置１０４は、処理装置１００とディスク装置１０５の間のデータ転送を行う。さらに、本発明においては、正制御装置１０４は、副制御装置１０９の間のデータ転送を行う機能をもつ。あるいは、正制御装置１０４が１つ以上のデイレクタ１０６を含み、各ディレクタ１０６が、処理装置２００とディスク装置２０５との転送、副制御装置１０９との間のデータ転送を行ってもよい。また、副制御装置１０９の内部構成は、正制御装置１０４と同様である。

【００１８】制御用メモリ１０７には、ライトデータ１１２に対応したライトデータ管理情報１１３が、作成される。

【００１９】処理装置１００は、正制御装置１０４にライト要求１１０を発行する時、ライトデータ１１２に、ライト時刻１１１に付与する。ライト時刻１１１は、本ライト要求１１０が発行された時刻を表しており、ライト時刻１１１により、処理装置１００が発行したライト要求１１０順序を認識することができる。処理装置１００が複数存在する場合、ライト時刻１１１は、処理装置１００間で、共通のクロックなどを用い、異なった処理装置１００で発行されたライト要求１１０の順序も、認識できるようになっているものとする。

【００２０】図２は、ライトデータ管理情報１１３の構成である。ここでは、特に、本発明に直接関係する情報について説明する。なお、本発明では、処理装置１００がライト要求１１０を発行する際、指定するディスクを論理ディスクとよぶ。論理ディスクＩＤ１２０は、対応するライトデータを書き込むよう、処理装置１００から指示された論理ディスクの番号であり、ライト要求１１０に含まれる情報である。本発明では、処理装置１００が認識している論理ディスクとディスク装置１０５（物理ディスク）は、１対１に対応している必要はない。図３にしめすように、論理ディスクが、複数のディスク装置１０５上に定義されてもよい。また、論理ディスクに、冗長データを含ませ、ＲＡＩＤ（Ｒｅｄｕｎｄａｎｔ Ａｒｒａｙ ｏｆ Ｉｎｅｘｐｅｓｉｖｅ Ｄｉｓ

ｋｓ）構成にしてもよい。ライトアドレス１２１は、対応するライトデータを書き込む論理ディスク内のアドレスを示す情報（例えば、論理ディスクの先頭から１ＭＢｙｔｅの領域というような情報）で、ライト要求１１０に含まれる情報である。ライトデータ長１２２は、対応するライトデータの長さを表す情報であり、ライト要求１１０に含まれる情報である。以上の情報は、いずれも、通常のライト要求１１０に含まれる情報である。ライトデータポインタ１２３は、対応するライトデータ１１２へのポインタである。ライト時刻１１１については、すでに、説明したとおりである。ライト要求１１０に、ライト時刻１１１を付与することが本発明の特徴の１つである。副制御装置転送必要ビット１２４は、副制御装置１０９に対応するライトデータ１１２の転送が必要であることを表す情報である。

【００２１】制御用メモリ１０８に含まれるもう１つの情報は、副論理ディスク番号１１４である。本情報は、正制御装置１０４の論理ディスク対応に存在する情報で、対応する論理ディスクの２重書きペアになっている副論理ディスクの番号、すなわち、副論理ディスクを格納している副制御装置１０９の番号と、副論理ディスクの副制御装置１０９内の論理ディスク番号を含む。もちろん、２重書きペアをもたない論理ディスクには、ヌル値が入るものとする。

【００２２】副制御装置１０９の制御用メモリ１０９にも、ライトデータ管理情報１１３が含まれる。

【００２３】フォーマットは、正制御装置１０４内のライトデータ管理情報１１３と同じでよい。ただし、副制御装置転送必要ビット１２４は、常にオフとなっているものとする。さらに、正論理ディスク番号１３１である。本情報は、副制御装置１０４の論理ディスク対応に存在する情報で、対応する論理ディスクの２重書きペアになっている正論理ディスクの番号、すなわち、正論理ディスクを格納している正制御装置１０４の番号と、正論理ディスクの正制御装置１０４内の論理ディスクア番号を含む。もちろん、２重書きペアをもたない論理ディスクには、ヌル値が入るものとする。

【００２４】正制御装置１０４の正ライトデータ受領部１３０は、処理装置１００から、ライト要求１１０を受け取ったとき、動作を開始する。まず、受け取ったライトデータ１１２を、キャッシュメモリ１０８に格納する。（ステッップ１３１）次に、正ライトデータ受領部１４０は、制御用メモリ１０８内のライトデータ管理情報１１３を、当該ライト要求対応に確保する。（ステッップ１３２）さらに、ライト要求に含まれるライト時刻１１１等の情報を確保したライトデータ管理情報１１３に格納し、ライトデータポインタ１２３、副制御装置転送必要ビット１２４の設定を行う。（ステッップ１３３）最後に、処理装置１００に、ライト要求１１０の完了報告を行う。（ステップ１３４）以上の処理には、ディスク装置１０５へのアクセスがないため、高速な応答が可能となる。ライトデータ１１２をディスク装置１０５に書き込む処理は、正制御装置１０４が後から実行する。この動作は、通常の制御装置の動作であるため、特に、詳細に記述しない。

【００２５】正制御装置１０４の正ライトデータ送信部１４０は、ライトデータ１１２を副制御装置１０９に送る機能をもつ。まず、副制御装置転送必要ビット１２４が設定されているライトデータ管理情報１１３の中で、ライト時刻が最も以前であるライトデータ１１３を、対応する副論理ディスク番号１３０を参照して、２重書きペアが存在する副制御装置１０９へ送る。ライトデータ１１２の長さ、書き込みを行う副論理ディスク内のアドレスは、ライトデータ管理情報１１３内の情報を参照して指定する。（ステップ１４１）次に、副制御装置１０９からの完了報告をまつ。（ステップ１４２）完了報告が返ってくると、副制御装置転送必要ビット１２４をオフする。（ステップ１４３）この後、ステップ１４０へ戻り、次に送信すべきライトデータを見つける。

【００２６】副制御装置１０９の副ライトデータ受領部１６０は、正制御装置１０４から、ライトデータ１１２を受け取った時動作する。副ライトデータ受領部１６０の処理内容は、ライトデータ管理情報１１３の設定において、副制御装置転送必要ビット１２４の設定を行わない（ステップ１６１）こと以外は、正ライトデータ受領部１４０の処理内容と同様である。

【００２７】副制御装置１０９の副ライトデータデステージ部１４０は、ライトデータ１１２をディスク装置１０５に書き込む機能をもつ。まず、ライトデータ管理情報１１３の中で、ライト時刻が最も以前である順にいくつかのライトデータ１１３を、ディスク装置１０５に書き込むことを決定し、しかるべき計算を行い、書き込みを行うディスク装置１０５と書き込みアドレスを決める。この計算方法は、通常のＲＡＩＤ等で用いられる方法であるため、詳細には記述しない。（ステップ１７１）次に、ライトデータ１１２をディス装置１０５に書き込むよう要求を複数並行して、ディスク装置１０５に発行する。（ステップ１７２）さらに、次に、ディスク装置１０５からの完了報告をまつ。（ステップ１７３）すべての要求の完了報告を受け取った後、ステップ１７０へ戻り、次にディスク装置１０５にデステージすべきライトデータ１１３を見つける。

【００２８】正制御装置１０４から副制御装置１０９へのライトデータ１１３の送信順序が、ライト時刻１１１の順番であるため、副制御装置１０９では、ある時刻を基準に、それ以前のライトデータ１１３はすべて保持でき、それ以降のライトデータ１１３はまったく保持しないという状態を作り出すことができる。これにより、正制御装置１０４が破壊されても、副制御装置１０９側で、トランザクションの等中結果を残さない回復処理が

可能となる。また、副制御装置１０９側で、ライトデータ１１３、ライト時刻１１２等の制御情報は、キャッシュメモリ１０７、制御用メモリ１１３などの不揮発性の半導体メモリに保持されるため、性能上のオーバヘッドは小さい。

【００２９】以上説明してきた内容は、正制御装置１０４から副制御装置１０９へのライトデータ１１２の転送がシリアライズされているため、十分な性能が得られない可能性がある。図４は、正制御装置１０４から副制御装置１０９へのライトデータ１１２の転送を並列に実行した場合の動作を表している。各処理部で、転送がシリアライズされている場合と変更があるのは、正ライトデータ送信部ａ３００、正基準時刻送信部１７０、副基準時刻受信部１８０、副ライトデータデステージ部ａ３１０と正障害時データ破棄部１９０である。以下、正ライトデータ送信部ａ３００の処理フローについて説明する。まず、副制御装置転送必要ビット１２４が設定されているライトデータ管理情報１１３の中で、ライト時刻が最も以前である順にいくつかのライトデータ１１３を、対応する副論理ディスク番号１３０を参照して、２重書きペアが存在する副制御装置１０９へそれぞれ並列に転送する。（ステップ３０１）次に、副制御装置１０９から、それぞれの完了報告が送られてくるのをまつ。（ステップ３０２）すべての完了報告が返ってくると、対応するライトデータ管理情報１１３の中の副制御装置転送必要ビット１２４をオフする。（ステップ３０３）この後、ステップ１５０へ戻り、次に送信すべきライトデータ１１２を見つける。

【００３０】ライトデータ１１２の転送を並列に実行すると、副制御装置１０９側で保持されるライトデータ１１２のライト時刻１１１の順序がくるう可能性がある。したがって、副制御装置１０９がデステージしてよいライトデータ１１２を決めるための判断基準となるライト時刻１１１を認識する必要がある。この場合、デステージシしてよいライトデータ１１２は、正制御装置１０４の中で、副制御装置転送必要ビット１２４がオンになっているライトデータ管理情報１１３の中で、最も以前のライト時刻１１１を基準時刻として、この基準時刻より以前のライト時刻１１１をもつライトデータ１１２ということになる。というのは、この基準時刻より以前のライト時刻１１１をもつライトデータ１１２はすべて、副制御装置１０９側に保持されているいることになるためである。一方、この基準時刻より後のライト時刻１１１をもつライトデータ１１２は、まだデステージしてはまずいライトデータ１１２であり、正制御装置１０４が破壊された場合、これらのライトデータ１１２はデステージせず、破棄する必要がある。

【００３１】正基準時刻送信部１７０は、副制御装置１０９に上述したデステージしてよい基準時刻を送信する機能をもつ。基準時刻は、上述したように、副制御装置転送必要ビット１２４がオンになっているライトデータ管理情報１１３の中で、最も以前のライト時刻１１１である。

【００３２】副基準時刻受信部１８０は、正制御装置１０４から受信した基準時刻を、デステージ許可時刻１８５として、制御用メモリ１０８に格納する。

【００３３】図４は、正制御装置１０４から副制御装置１０９へのライトデータ１１２の転送を並列に実行した場合の副ライトデータデステージ部ａ３１０の処理フローである。図１に示した処理フローとは異なるのは、デステージするライトデータ１１３を選択する条件に、デステージ許可時刻１８５より以前のライト時刻１１１であるかどうかという条件が入ることだけである（ステップ３１１）。

【００３４】正障害時データ破棄部１９７は、正制御装置１０４が破壊された時、デステージ許可時刻１８５から後のライト時刻１１１をもつライトデータ１１２を破棄する機能をもつ（ステップ１９１）。

【００３５】次に、第２の実施例について説明する。

【００３６】図５は、第２の実施例の概要を表す。第２の実施例と第１の実施例の相違は、正制御装置１０４と副制御装置１０９の数である。第１の実施例では、正制御装置１０４の数が１台で、副制御装置１０９の数が１台であった。一方、第２の実施例では、正制御装置１０４の数が２台以上で、副制御装置２０９の数が１台である。

【００３７】正制御装置１０４が複数存在すると、副制御装置１０９側で、それぞれの正制御装置１０４から受け取っているライトデータ１１２のライト時刻１１１にずれが生ずる。一方の正制御装置１０４（例えば、正制御装置ａ）から受け取った最も最近のライト時刻１１１、この時刻を時刻ａ、もう一方の正制御装置１０４（例えば、正制御装置ｂ）から受け取った最も最近のライト時刻１１１、この時刻を時刻ｂとする。この場合、時刻ａより、時刻ｂの方が、以前の時刻であるとすると、正制御装置ａ側に、時刻ａより最近で、時刻ｂより以前のライトデータ１１３を保持している可能性がある。すでにのべたように、トランザクションの途中結果を残さないようにするには、ある基準時刻以前のライト時刻１１２をもつライトデータ１１３はすべて保証し、基準時刻以降のライト時刻１１２をもつライトデータ１１３はすべて破棄する必要がある。したがって、時刻ａ以前のライト時刻１１１をもったライトデータ１１２が、副制御装置１０９でデステージしてよいライトデータ１１２ということになる。

【００３８】以上に対応して、副制御装置１０９の制御用メモリ１０８には、正制御装置ライト許可時刻５００がある。正制御装置ライト許可時刻５００は正制御装置１０４ごとに存在する情報で、対応する正制御装置１０４から受け取った最も最近のライト時刻１１１が格納さ

れている。したがって、上述したように、これらの正制御装置ライト許可時刻５００の中で、もっとも以前の時刻を基準時刻として、この基準時刻以前のライト時刻１１１をもったライトデータ１１２が、副制御装置１０９でデステージしてよいライトデータ１１２ということになる。

【００３９】以下、本実施例でも、１台の正制御装置１０４から副制御装置１０９へのライトデータ１１２の転送を並列に実行した場合の各処理部の内容について述べる。もちろん、１台の正制御装置１０４から副制御装置１０９へのライトデータ１１２の転送をシリアライズして実行した場合についても、本実施例は有効である。

【００４０】正制御装置１０４の各処理部の処理フローは第１の実施例で、ライトデータ１１２の転送を並列に実行した場合（図３の処理）の処理フローと同様である。

【００４１】次に、副制御装置１０９の各処理部の処理フローの説明を行う。

【００４２】第２の実施例における副ライトデータデステージ部ｂ５１０の処理フローについて説明する。ここでは、第２の実施例における副ライトデータデステージ部５２０の処理フローが、第１の実施例における副ライトデータデステージ部１７０の処理フローと異なる点について説明する。第２の実施例における副ライトデータ受領部５１０の処理内容は、デステージするライトデータ１１２を選択する際、対応するライト時刻１１１が、すべての正制御装置ライト許可時刻５００より以前であるかをチェックして、条件を満たすライトデータ１１２を選択する点である。（ステップ５１１）これ以外は、第２の実施例における副ライトデータデステージ部ｂ５１０の処理フローは、第１の実施例における副ライトデータデステージ部１７０の処理フローと同様である。

【００４３】副基準時刻受信部ｂ５２０は、正制御装置１０４から受信した基準時刻を、基準時刻を送信してきた正制御装置１０４に対応する正制御装置ライト許可時刻５００に設定する。

【００４４】本実施例においては、正障害時データ破棄部ｂ５３０が、破棄するライトデータ１１２は、対応するライト時刻１１１が、すべての正制御装置ライト許可時刻５００より以前であるという条件を満足しないライトデータ１１２である。（ステップ５３１）次に、第３の実施例について説明する。

【００４５】図６は、第の実施例の概要を表す。第３の実施例と第２の実施例の相違も、正制御装置１０４と副制御装置１０９の数である。第３の実施例では、正制御装置１０４の数が２台以上で、副制御装置１０９の数は１台以上である。この場合、すべての正制御装置１０４と副制御装置１０９のペアがお互いに、接続されている必要はない。

【００４６】副制御装置１０９が複数存在すると、トランザクションの途中結果を残さないようにするには、各副制御装置１０９間で、デステージするライトデータ１１２を選択する際に用いる基準時刻を共通にする必要がある。これは、データベースやジャーナルが複数の副制御装置１０９間に分散している可能性があるためである。

【００４７】本実施例では、デステージするライトデータ１１２を選択する際に用いる基準時刻を決定する機能を、マスタ副制御装置７００に持たせる。したがって、マスタ制御装置７００とそれ以外の副制御装置１０９の間は、データ転送路で接続されている。データ転送路が故障すると、各副制御装置１０９間で、デステージするライトデータ１１２を選択する際に用いる基準時刻を共通化することができなくなるため、データ転送路は多重化しておくことが望ましい。本実施例では、デステージするライトデータ１１２を選択する際に用いる基準時刻を決定する機能を、マスタ副制御装置７００に持たせたが、基準時刻を決定する機能を、特定の副制御装置１０９に持たせず、各副制御装置１０９に分散させる方法（例えば、交代で、各副制御装置１０９が基準時刻を決定するような方法）をとっても、本発明は有効である。

【００４８】以上に対応して、マスタ副制御装置７００の制御用メモリ１０８には、副制御装置ライト時刻７０１がある。副制御装置ライト時刻７０１は、マスタ副制御装置７００も含めた副制御装置１０９対応の情報である。各副制御装置ライト時刻７０１は、各副制御装置１０９から、マスタ副制御装置７００が、適当な周期で、その副制御装置１０９内のすべての正制御装置ライト許可時刻５００の中で、もっとも以前の時刻（実施例２で、副制御装置１０９がライトデータの選択の際、基準とした時刻）受け取る情報である。

【００４９】マスタライト時刻７０２は、第３の実施例において、各副制御装置１０９がライトデータの選択の際、基準とする時刻である。マスタライト時刻７０２は、マスタ副制御装置７００が、適当な周期で、すべての副制御装置ライト時刻７０１を参照して、もっとも以前の時刻を選択して、この時刻を設定する。選択した時刻以前のライト時刻１１１をもったすべてのライトデータ１１２は副制御装置１０９に保持されていることになる。このため、この条件を満足するライトデータ１１２を保証して、満足しないライトデータ１１２はすべて破棄することにより、トランザクションの途中結果を残さないようにすることができる。

【００５０】以下、本実施例でも、１台の正制御装置１０４から副制御装置１０９へのライトデータ１１２の転送を並列に実行した場合の各処理部の内容について述べる。もちろん、１台の正制御装置１０４から副制御装置１０９へのライトデータ１１２の転送をシリアライズして実行した場合についても、本実施例は有効である。

【００５１】正制御装置１０４の各処理部の処理フロー

は第２の実施例とほとんど同様である。もちろん、正基準時刻送信部１７０はライトデータ１１２を送信する各副制御装置１０９に基準時刻を送る機能をもつ。ただし、送信する基準時刻は、送信を行う副制御装置１０９に対応するすべてのライトデータ１１２の副制御装置転送必要ビット１２４がオンになっているライトデータ管理情報１１３の中で、最も以前のライト時刻１１１である。

【００５２】正制御装置１０４の各処理部の処理フローの中で、第２の実施例と異なるのは、副制御装置１０９に、副ライト時刻送信部７１０、副ライトデータデステージ部ｃ７２０、正障害時データ破棄部ｃを含む点である。また、マスタ副制御装置７００は、マスタ副ライト時刻受信部７１１、マスタライト時刻計算部７１２、マスタ副ライト時刻送信部７１３７３０を含む点である。

【００５３】副ライト時刻送信部７１０は、適当な周期で、その副制御装置１０９内のすべての正制御装置ライト許可時刻５００の中で、もっとも以前の時刻１８０を、マスタ副制御装置７００のマスタ副ライト時刻受信部７１１に送る。マスタ副制御装置７００以外の副制御装置１０９の副ライト時刻送信部７１０は、副制御装置１０９間のデータ転送路を利用する。マスタ副制御装置７００の副ライト時刻送信部７１０は、マスタ副制御装置７００の通信手段を利用する。

【００５４】マスタ副ライト時刻受信部７１１は、副ライト時刻送信部７１０から受信した時刻を、当該時刻を送ってきた副制御装置１０９に対応する副制御装置ライト時刻７０１に、設定する。

【００５５】マスタライト時刻計算部７１２は、適当な周期で、すべての副制御装置ライト時刻７０１を参照して、もっとも以前の時刻を選択して、この時刻を、マスタライト時刻７０２に設定する。

【００５６】マスタ副ライト時刻送信部７１３は、各副制御装置１０９の副ライトデータデステージ部７２０、正障害時データ破棄部７３０からの要求にしたがって、適当な周期で、マスタライト時刻７０２に設定された時刻を送る。マスタ副制御装置７００以外の副制御装置１０９には、副制御装置１０９間のデータ転送路を利用する。マスタ副制御装置７００からの要求には、マスタ副ライト時刻送信部７１３は、マスタ副制御装置７００の通信手段を利用する。

【００５７】副ライトデータデステージ部７２０が、第２の実施例と異なる点は、デステージを行うライトデータ１１２を選択する際、マスタ副ライト時刻送信部７１３から基準となる時刻を受信し、この時刻より以前のライト時刻１１１をもつライトデータ１１２をデステージの対象として選択する点である（ステップ７２１）。

【００５８】正障害時データ破棄部ｃ７３０が、第２の実施例と異なる点は、キャッシュメモリ１０７から破棄するライトデータ１１２を選択する際、マスタ副ライト時刻送信部７１３から基準となる時刻を受信し、この時刻より以前のライト時刻１１１をもつライトデータ１１２以外のライトデータ１１２を破棄の対象として選択する点である。（ステップ７３１）本実施例では、マスタ制御装置７００が、副制御装置１０９から基準となる時刻を計算するのに必要な情報を受け取ったが、図７に示すように、正制御装置１０４から受け取るようにしてもよい。この場合、マスタ副制御装置７００の制御用メモリ１０８には、マスタ正制御装置ライト時刻８００がある。マスタ正制御装置ライト時刻８００は、正制御装置１０４対応の情報である。マスタ正制御装置ライト時刻８００は、各正制御装置１０４から、マスタ副制御装置７００が、適当な周期で、その正制御装置１０９内で、オン状態の副制御装置転送必要ビット１２４を含むすべてのライトデータ管理情報１１３の中で、最も以前のライト時刻１１１を受け取り、設定を行う情報である。マスタライト時刻７０２には、マスタ副制御装置７００が、適当な周期で、すべての正制御装置ライト時刻７０１を参照して、もっとも以前の時刻を選択して、この時刻を設定する。デステージ、データ破棄の際、基準となる時刻として用いる時刻が、マスタライト時刻７０２である点は、同様である。

【００５９】

【発明の効果】本発明の目的は、遠隔地の制御装置間で２重書きを行う際、直接制御装置間で、ライトデータを受け渡すことにより、データ転送数を必要最小限に留め、しかも、制御装置間の距離が拡大しても、性能の劣化を微小に押さえ、さらに、トランザクションの途中結果を残さないような機能を提供することである。加えて、制御情報のディスク入出力処理の実行も不必要とし、高性能化を図る。

【図面の簡単な説明】

【図１】第１の実施例の概要図。

【図２】ライトデータ管理情報のフォーマット。

【図３】正制御装置から副制御装置へのライトデータの転送を並列に実行した場合の正ライトデータ送信部の処理フロー図。

【図４】正制御装置から副制御装置へのライトデータの転送を並列に実行した場合の副ライトデータデステージ部の処理フロー図。

【図５】第２の実施例の概要図。

【図６】第３の実施例の概要図。

【図７】第３の実施例において、マスタ副制御装置が、正制御装置からライト時刻に関する情報を収集した場合の処理概要図。

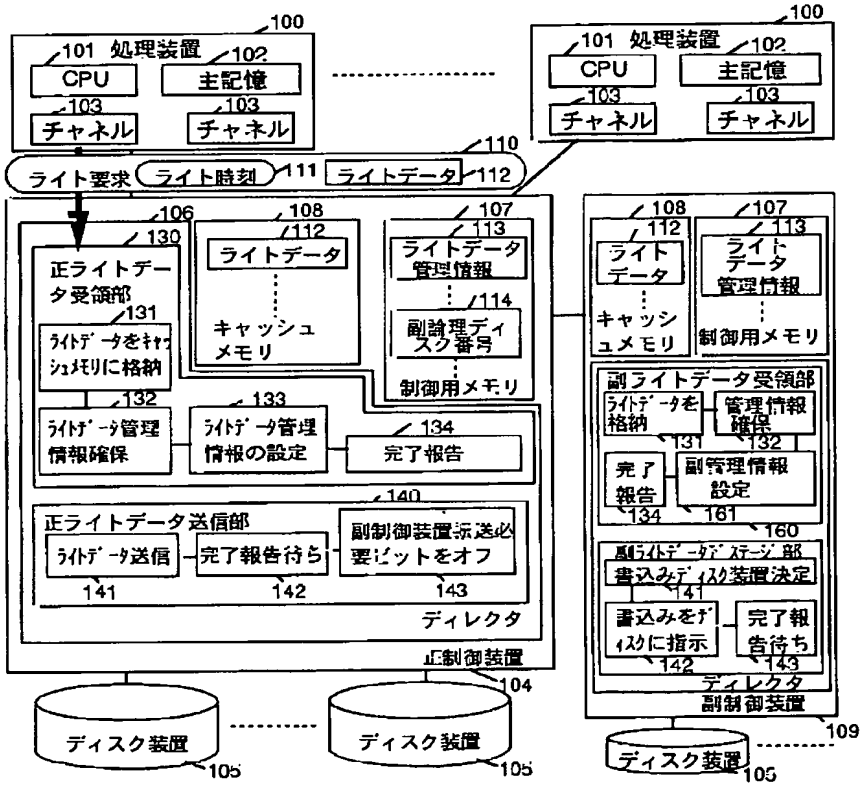
【符号の説明】

１０４…正制御装置、１０７…制御用メモリ、１０８…キャッシュメモリ、１０９…副制御装置、１１１…ライト時刻、１１２…ライトデータ、１１３…ライトデータ管理情報、１２４…副制御装置転送必要ビット、１３０

…正ライトデータ受領部、１４０…正ライトデータ送信部、１５０…副ライトデータ受領部、１６０…副ライトデータデステージ部、１７０…正基準時刻送信部、１８０…副基準時刻受信部、１８５…デステージ許可時刻、１９０…正障害時データ破棄部、３００…正ライトデータ送信部ａ、３１０…副ライトデータデステージ部ａ、５００…正制御装置ライト許可時刻、５１０…副ライトデータデステージ部ｂ、５２０…副基準時刻受信部ｂ、５３０…正障害時データ破棄部ｂ、７００…マスタ副制御装置、７０１…副制御装置ライト時刻、７０２…マスタライト基準時刻、７１０…副ライト時刻送信部、７１１…マスタ副ライト時刻受信部、７１２…マスタ副ライト時刻ケ計算部、７１３…マスタ副ライト時刻送信部、７１０…副ライトデータデステージ部ｃ、７２０…正障害時データ破棄部ｃ、８００…正制御装置ライト時刻。

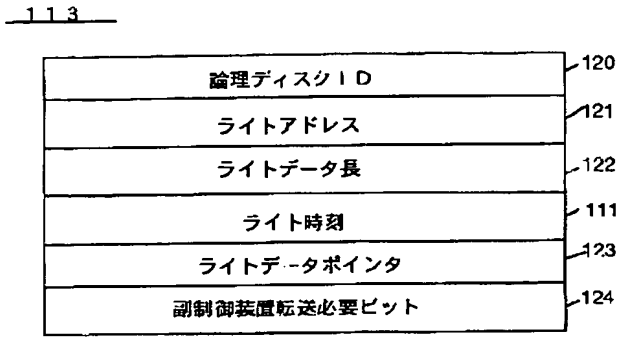
【図１】

図１



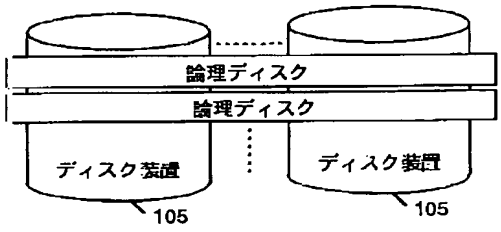
【図２】

図２



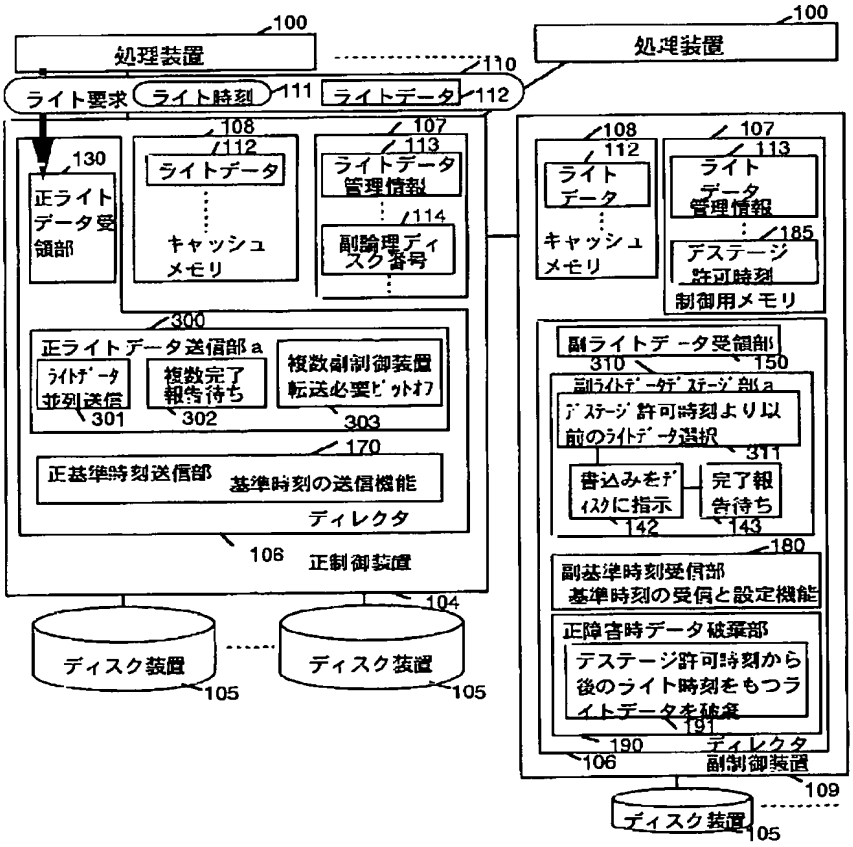
【図３】

図３



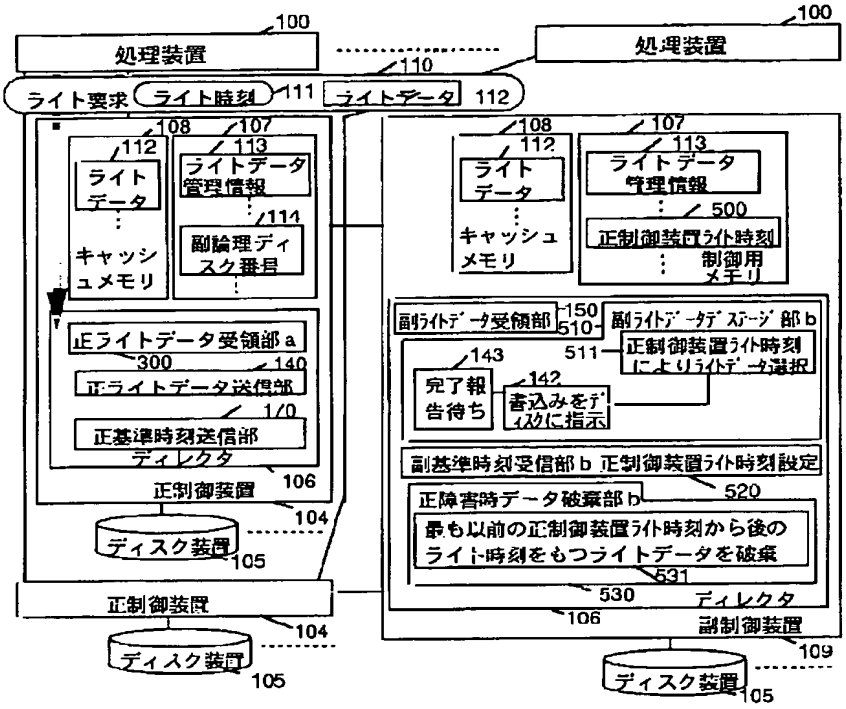
【図４】

図４



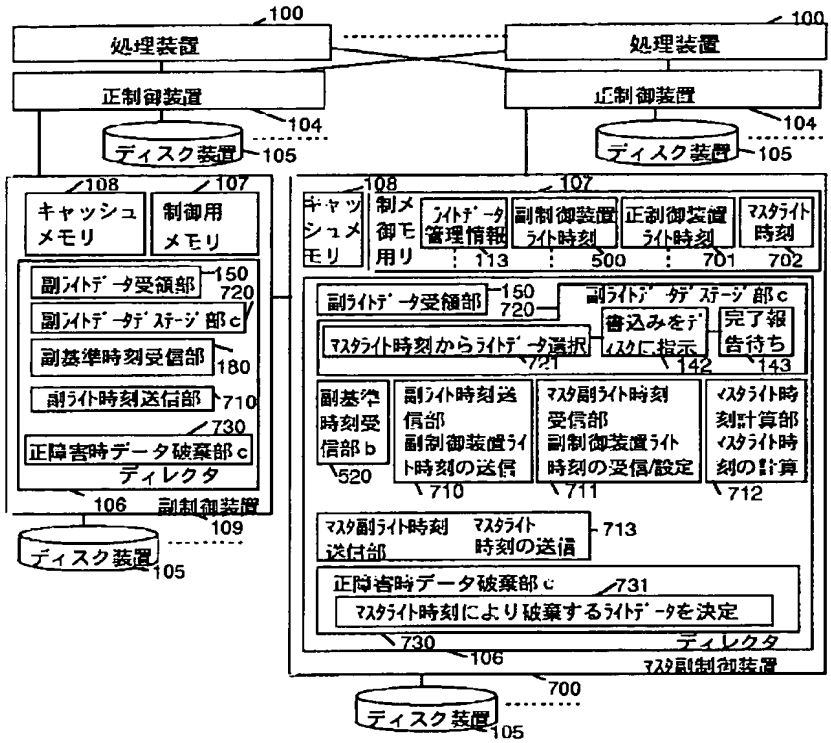
【図５】

図５


100
処理装置
110
ライト要求
ライト時刻
111
ライトデータ
112
108
107
112
113
ライトデータ管理情報
114
副論理ディスク番号
キャッシュメモリ
正ライトデータ受領部 a
300
140
正ライトデータ送信部
170
正基準時刻送信部
ディレクタ
106
正制御装置
104
ディスク装置
105
副ライトデータ受領部
150
510
副ライトデータステージ部 b
511
正制御装置ライト時刻によりライトデータ選択
143
完了報告待ち
142
書込みをディスクに指示
副基準時刻受信部 b 正制御装置ライト時刻設定
520
正障害時データ破棄部 b
最も以前の正制御装置ライト時刻から後のライト時刻をもつライトデータを破棄
531
530
ディレクタ
106
副制御装置
109
500
正制御装置ライト時刻
制御用メモリ

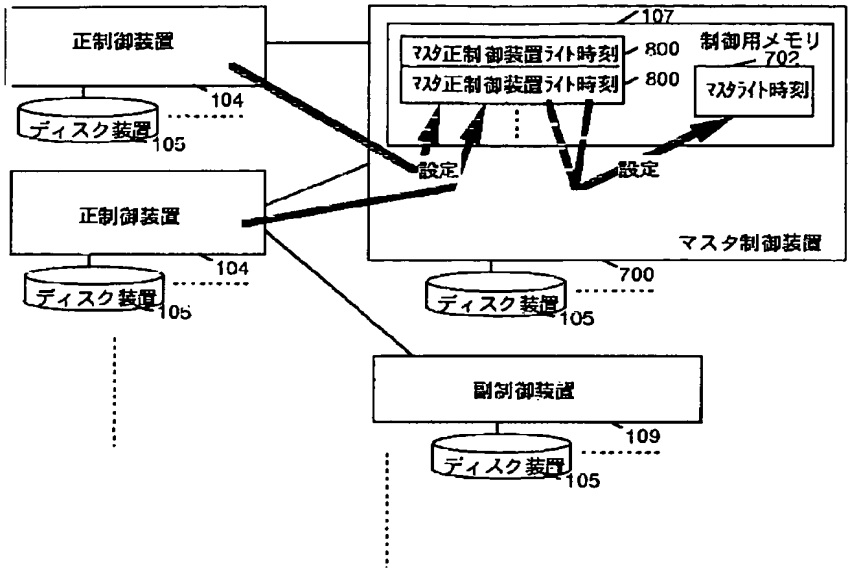
【図６】

図６


100
処理装置
正制御装置
104
ディスク装置
105
108
キャッシュメモリ
107
制御用メモリ
副ライトデータ受領部
150
720
副ライトデータステージ部c
副基準時刻受信部
180
副ライト時刻送信部
710
730
正障害時データ破棄部c
ディレクタ
106
副制御装置
109
ディスク装置
105
100
処理装置
正制御装置
104
ディスク装置
105
108
キャッシュメモリ
制御用メモリ
ライトデータ管理情報
113
副制御装置ライト時刻
500
正制御装置ライト時刻
701
マスタライト時刻
702
副ライトデータ受領部
150
720
副ライトデータステージ部c
マスタライト時刻からライトデータ選択
721
書込みをディスクに指示
142
完了報告待ち
143
副基準時刻受信部b
520
副ライト時刻送信部 副制御装置ライト時刻の送信
710
マスタ副ライト時刻受信部 副制御装置ライト時刻の受信/設定
711
マスタライト時刻計算部 マスタライト時刻の計算
712
マスタ副ライト時刻送付部 マスタライト時刻の送信
713
正障害時データ破棄部c
731
マスタライト時刻により破棄するライトデータを決定
730
106
ディレクタ
マスタ副制御装置
700
ディスク装置
105

【図７】

図７


正制御装置
104
ディスク装置
105
正制御装置
104
ディスク装置
105
107
マスタ正制御装置ライト時刻
800
マスタ正制御装置ライト時刻
800
制御用メモリ
702
マスタライト時刻
設定
設定
マスタ制御装置
700
ディスク装置
105
副制御装置
109
ディスク装置
105